# Measuring Icebergs @CoronaSurveys

Carlos Baquero & @CoronaSurveys Team

HASLab, INESC TEC & Univ. Minho, Portugal

Infoblender, Braga, 24 April 2020


Universidade do Minho

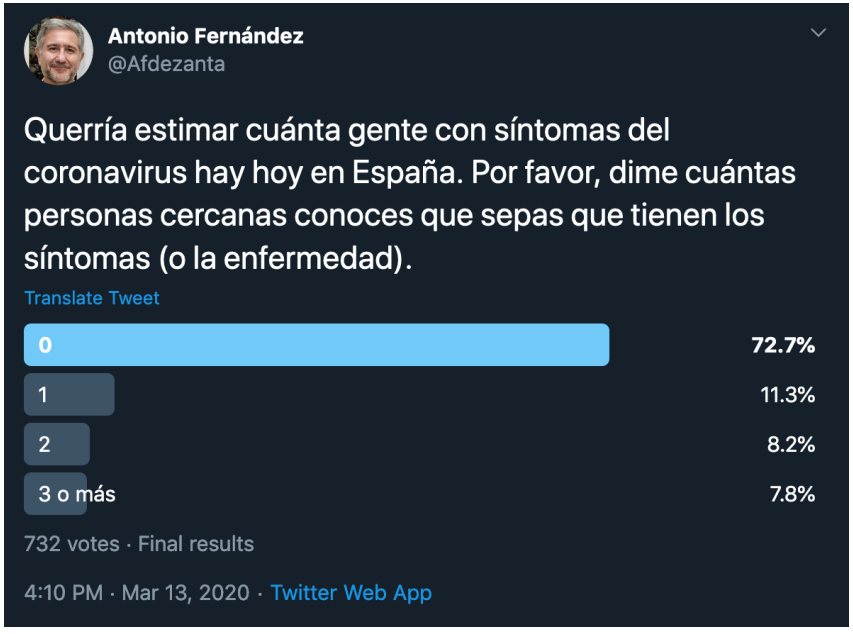
# Spain, March 13th 2020

- El País "Spain is the second EU country in number of cases"
- Reuters "The Spanish health ministry said the number of coronavirus cases in the country jumped to 4,209 on Friday from Thursday's 3,004 as the disease spread mostly in Madrid, the Basque Country and La Rioja regions."

Under rapid growth it is hard to keep up with testing and there are delays in reporting. This was the case in Spain.

# Open Surveys via Twitter

Antonio Fernandez Anta, IMDEA Networks, Spain




- $cases = 374,\ reach = 732 * 150$ (Dunbar number is 150)
- $total = \frac{cases}{reach} * ESpop \approx 153000$ ($\approx 30*$ official)

# Trade-offs in Open Surveys

Negative Points:

- Unorthodox approach: no panel selection, no stratification
- Not possible to track answers from users across time
- Fault injections

Positive Points:

- No personal data: health status of respondent is not asked
- No GDPR issues (but Ethics Board approval)
- Number of answers and reach size

# @CoronaSurveys at http://coronasurveys.org

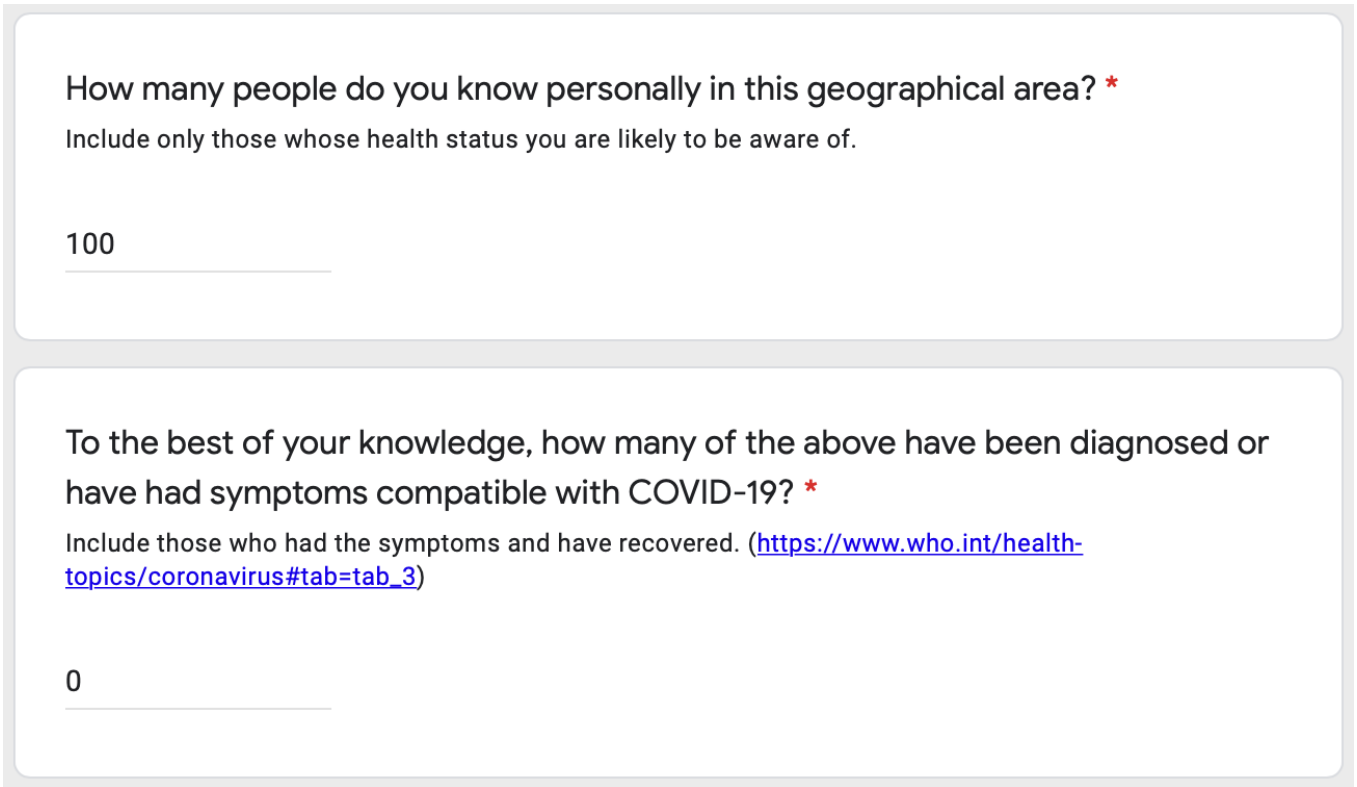
**@CoronaSurveys: Monitoring the Incidence of COVID-19 via Open Surveys**

## CoronaSurveys Project Summary

The CoronaSurveys project is a collaborative endeavour from several universities and research institutions (team members). Data about COVID-19 cases is collected via anonymous open surveys (all the data collected is openly available). The results below present estimations on the incidence of COVID-19 from this and other available data. You can help by regularly completing the anonymous survey.

# Two “simple” questions

For any given country and after user informed consent we ask:


How many people do you know personally in this geographical area? *

Include only those whose health status you are likely to be aware of.

100

To the best of your knowledge, how many of the above have been diagnosed or have had symptoms compatible with COVID-19? *

Include those who had the symptoms and have recovered. (https://www.who.int/health-topics/coronavirus#tab=tab_3)

0

# Localization: Ukraine

Скількох людей ви знаєте особисто в цій географічній зоні? *

Включіть лише тих, про стан здоров'я яких вам відомо особисто, не зі ЗМІ.

100

За вашими даними, скільком людям із зазначених вище було діагностовано COVID-19, або скільки мали симптоми, схожі з COVID-19? *

Включіть тих, у кого були симптоми та хто одужав. (Дізнайтесь про симптоми: https://www.who.int/health-topics/coronavirus#tab=tab_3)

0

# Data cleaning

Survey responses, pairs $(cases_i, reach_i)$, are cleaned by identifying and removing outliers:

- $reach_i$ – remove entries above $1.5*$ the interquartile range
- $\frac{cases_i}{reach_i} < 0.3$ – remove entries with very high incidence

We are migrating to a combined outlier detection method

# Estimators

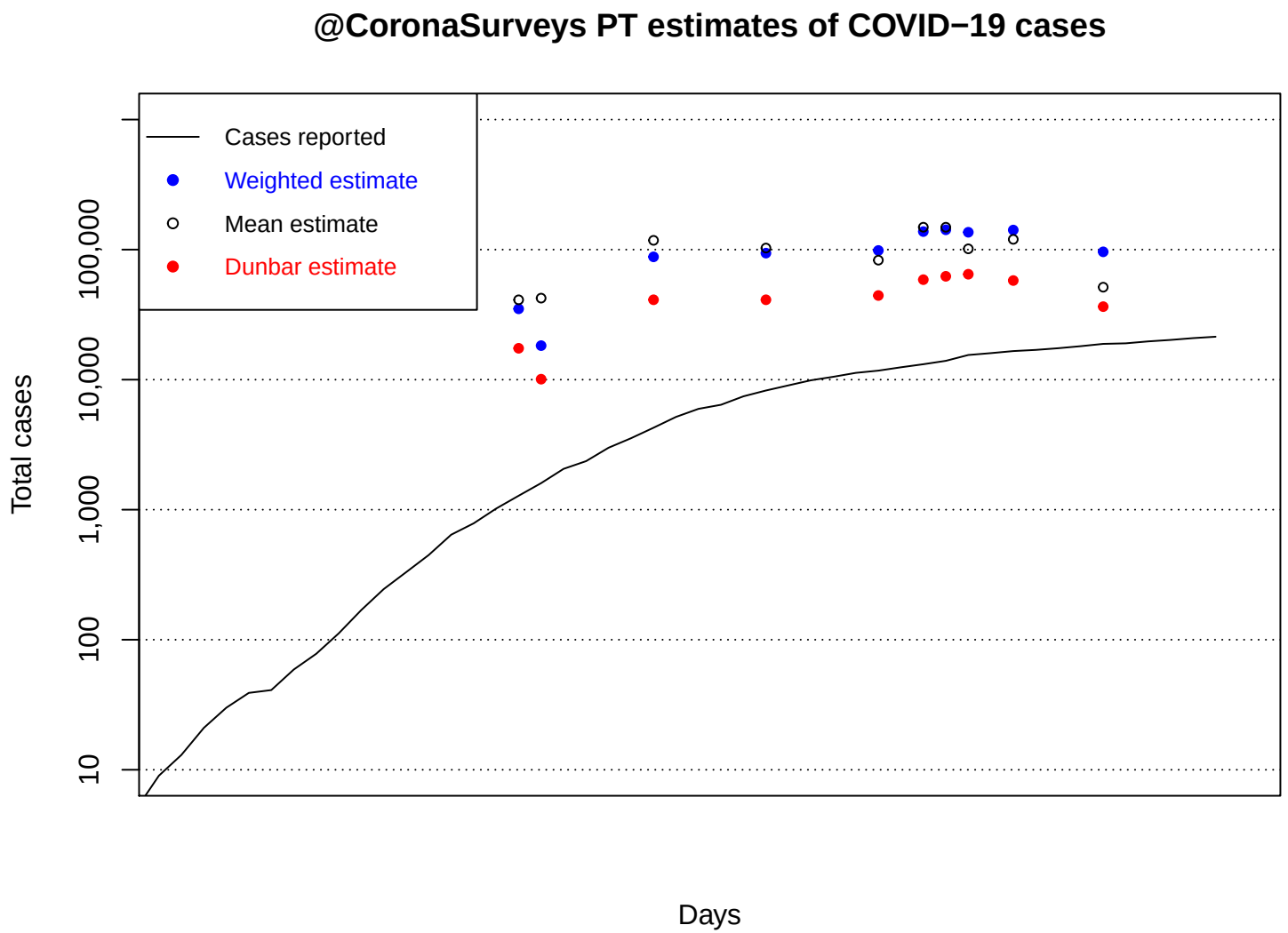
Assume $n$ response pairs $(cases_i, reach_i)$, region of population $P$
We can produce three estimators:

$$\text{Weighted} \quad E_w = P \cdot \frac{\sum_i cases_i}{\sum_i reach_i}$$

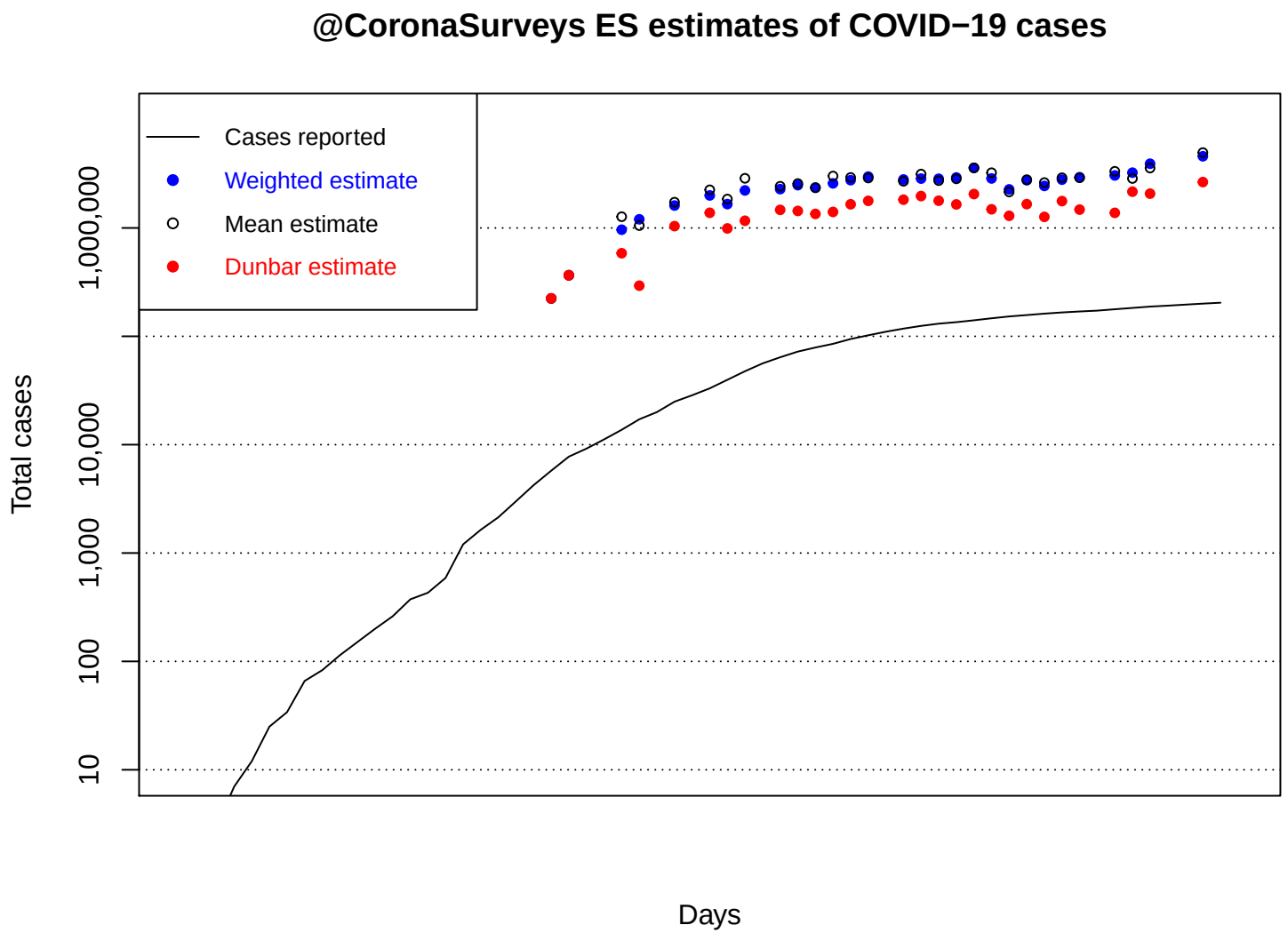
$$\text{Mean} \quad E_m = P \cdot \frac{\sum_i (cases_i / reach_i)}{n}$$

$$\text{Dunbar} \quad E_d = P \cdot \frac{\sum_i cases_i}{n \cdot 150}$$

# Estimates, Portugal



# Estimates, Spain



# Discussion

- Several datapoints, from batches of 30 or more answers
- Consistent evolution across time

- Still, how can we know the data is meaningful?
- We need other independent estimates

# Discussion

- Several datapoints, from batches of 30 or more answers
- Consistent evolution across time

- Still, how can we know the data is meaningful?
- We need other independent estimates

# Fatality based estimates

## Coarse grained: deaths * 400 estimate

Amy Maxmen. How much is coronavirus spreading under radar? Nature News Explainer, March 13th, 2020. https://www.nature.com/articles/d41586-020-00760-8

### Current deaths times 400

> goes something like this: Data from China suggest that about three weeks passes between when a person feels sick and dies from COVID-19. And if you assume a case fatality rate of roughly 1%, a back-of-the-envelope calculation suggests that each death represents about 100 cases in the first week. Right now, he adds, you can expect the epidemic to double each week if those cases aren't identified and isolated – bringing the estimate to 400 at the time of death. Because the error bars on each of these variables are large, epidemiologists check their figures against further information.

Coarse grained estimate, only in the initial exponential growth.

# Fatality based estimates

## Fine grained: cCFR estimate

- Corrected Case Fatality Ratio $\approx$ fatalities over detected cases with known outcomes (about 2 weeks lag)
- Find a "stable" baseline for the cCFR
  - China, Wuhan, $cCFR = 1.38$. Verity et al. Lancet paper

# Fatality based estimates

## Fine grained: cCFR estimate

- Corrected Case Fatality Ratio $\approx$ fatalities over detected cases with known outcomes (about 2 weeks lag)
- Find a "stable" baseline for the cCFR
  - China, Wuhan, $cCFR = 1.38$. Verity et al. Lancet Paper

### April 17th, Wall Street Journal

# Fatality based estimates

## Fine grained: cCFR estimate

- Corrected Case Fatality Ratio $\approx$ fatalities over detected cases with known outcomes (about 2 weeks lag)
- Find a "stable" baseline for the cCFR
  - China, Wuhan, $cCFR = 1.38$. Verity et al. Lancet Paper
  - Alternatives: Korean data, Germany Gangelt data
- Calculate a country cCFR from daily ECDC data
- Use proportion to baseline to infer current cases estimate

Caveats: Baseline quality, death reporting policies, cases coverage, calibration of symptoms to death delay, missing date of symptoms.

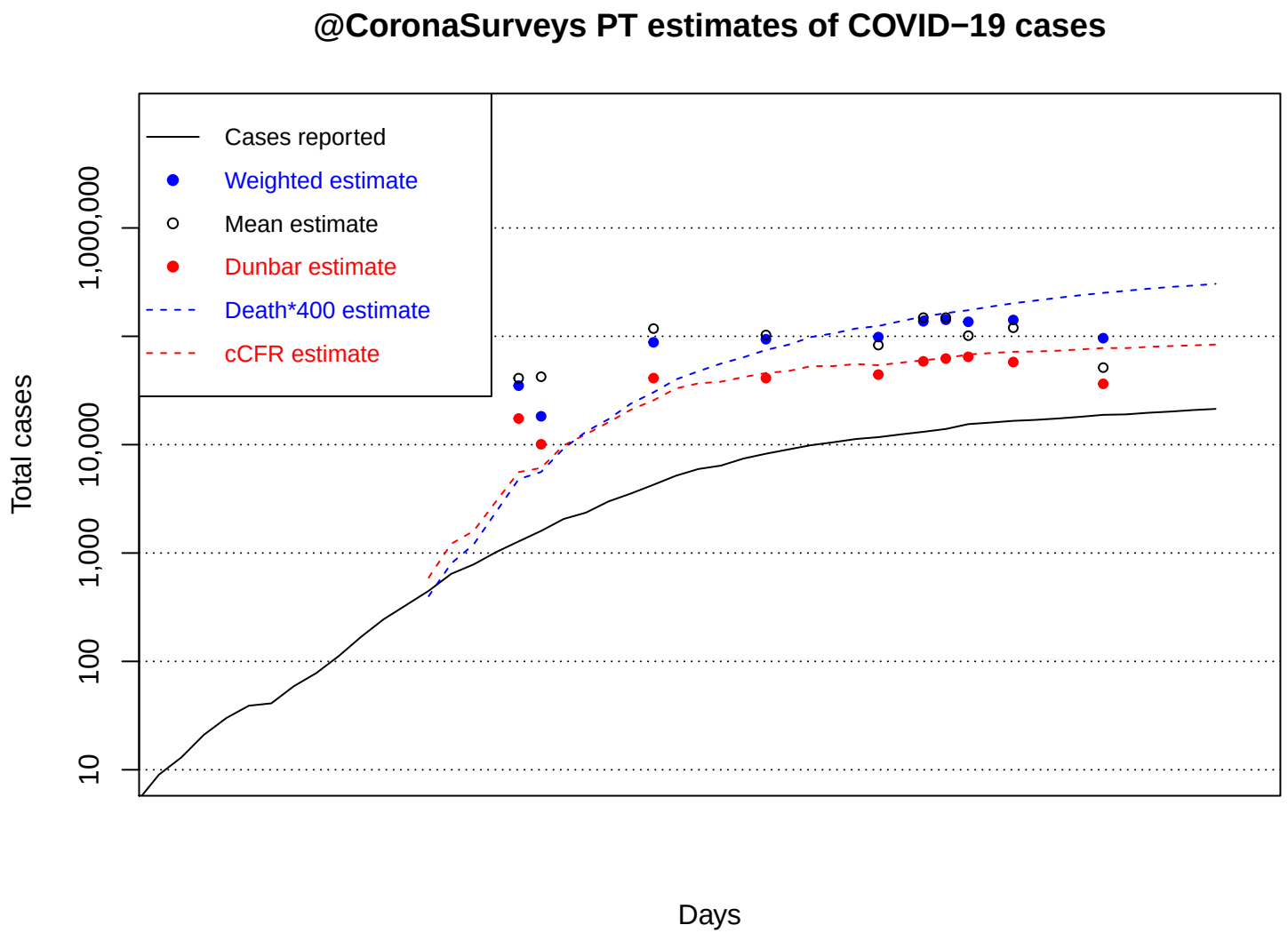
# Fatality based estimates

## Fine grained: cCFR estimate

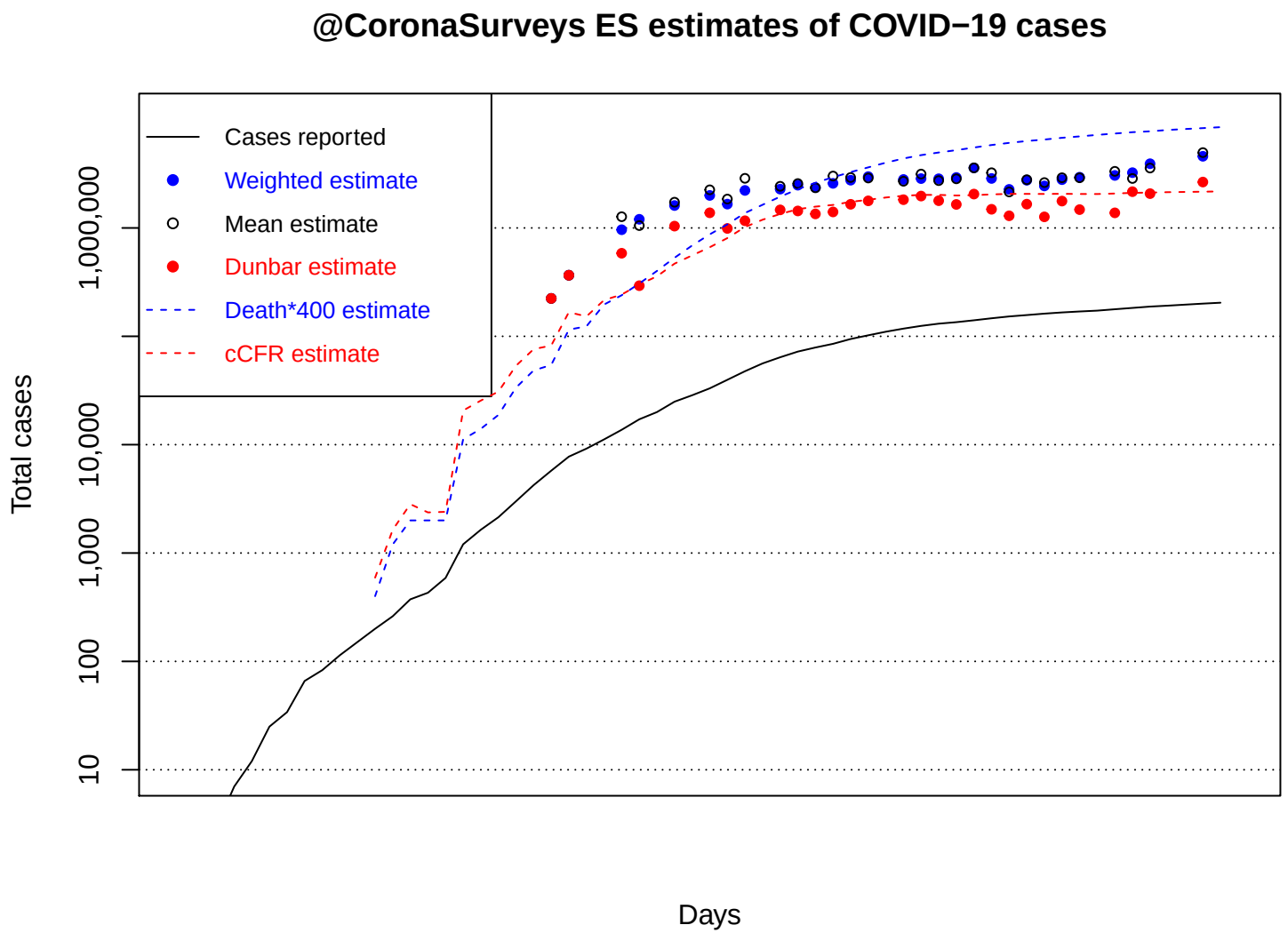
- Corrected Case Fatality Ratio $\approx$ fatalities over detected cases with known outcomes (about 2 weeks lag)
- Find a "stable" baseline for the cCFR
  - China, Wuhan, $cCFR = 1.38$. Verity et al. Lancet Paper
  - Alternatives: Korean data, Germany Gangelt data
- Calculate a country cCFR from daily ECDC data
- Use proportion to baseline to infer current cases estimate

Caveats: Baseline quality, death reporting policies, cases coverage, calibration of symptoms to death delay, missing date of symptoms.

# Estimates, Portugal



# Estimates, Spain



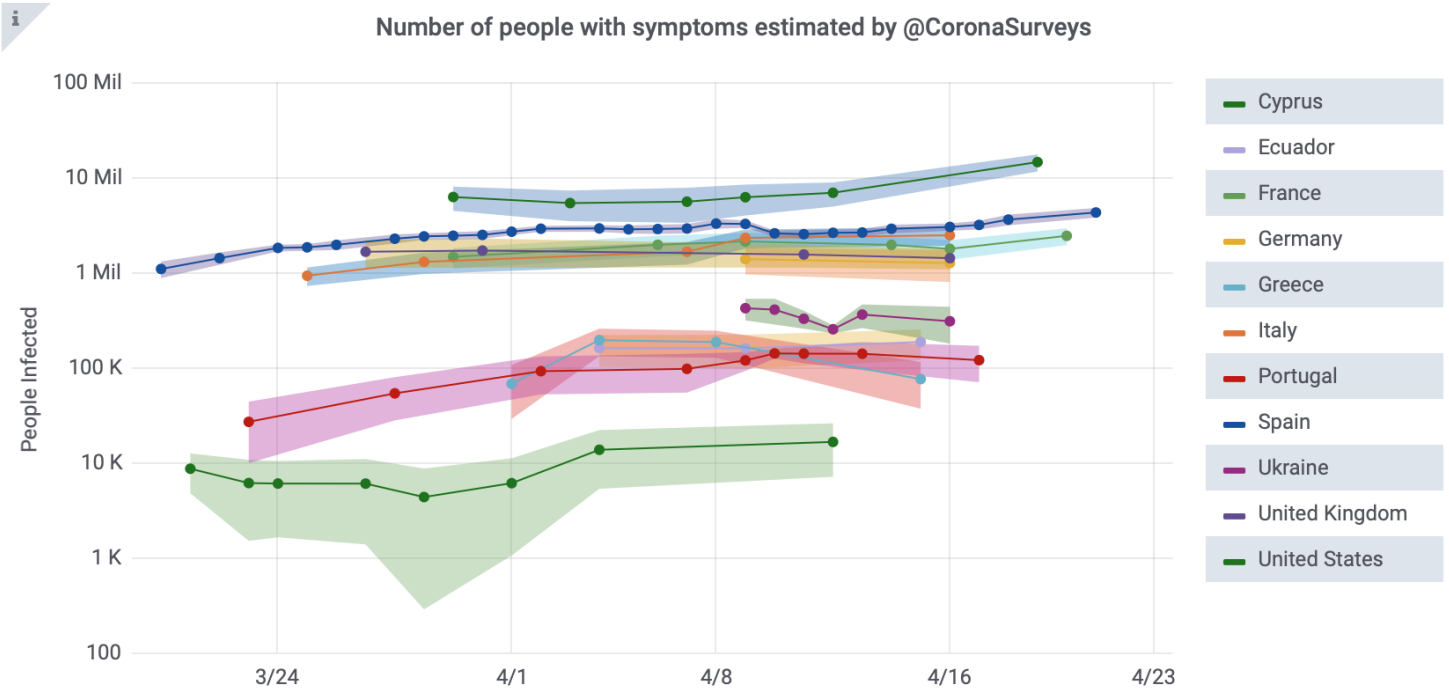
# @Coronasurveys, status and potential uses

- Data is being collected since mid March
- We are calibrating as the pandemic and data evolves
- Surveys are open for all globe

- Potential for quick assessment when reliable techniques lack
- Countries with good digital penetration but lacking testing

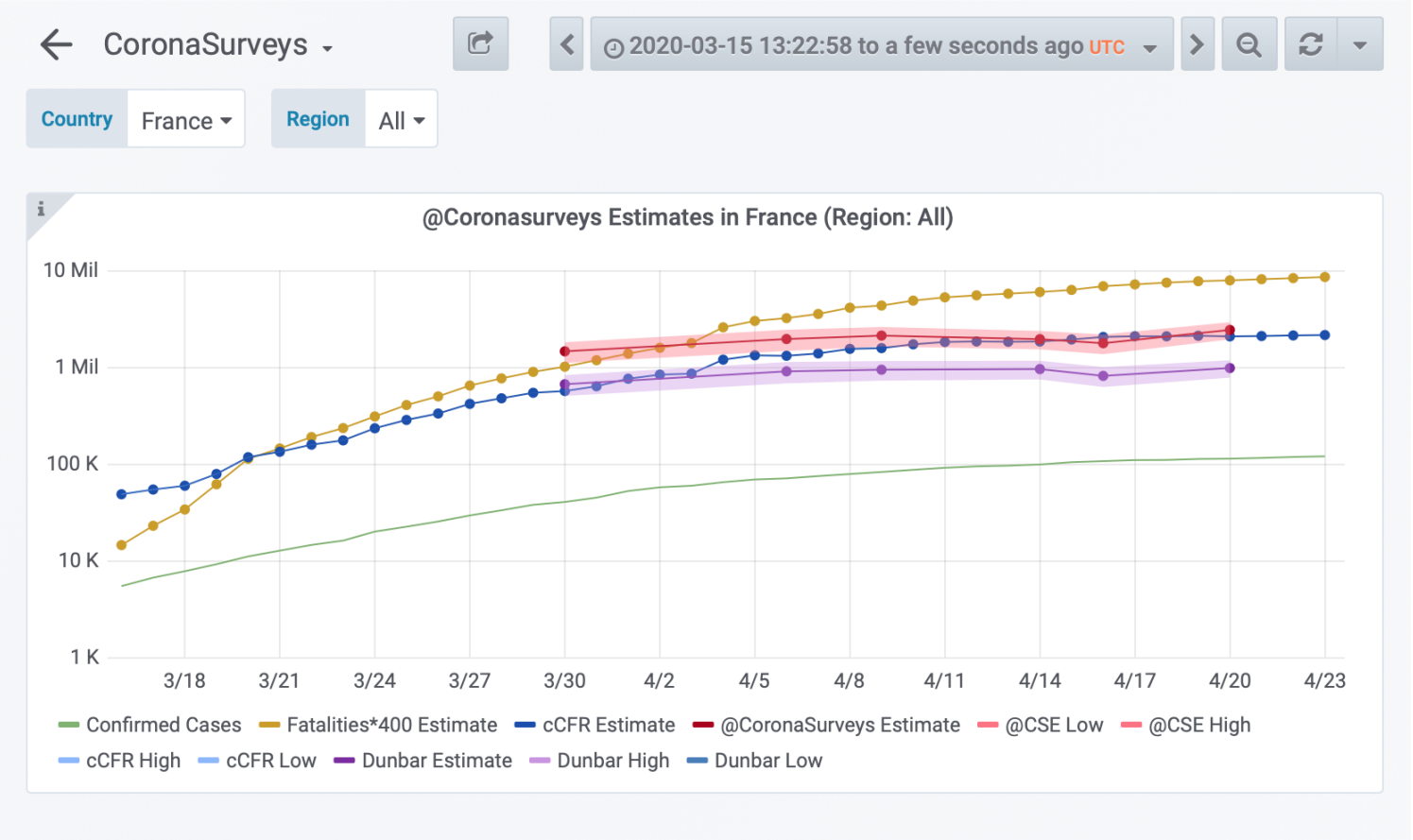
# Portfolio

## Countries

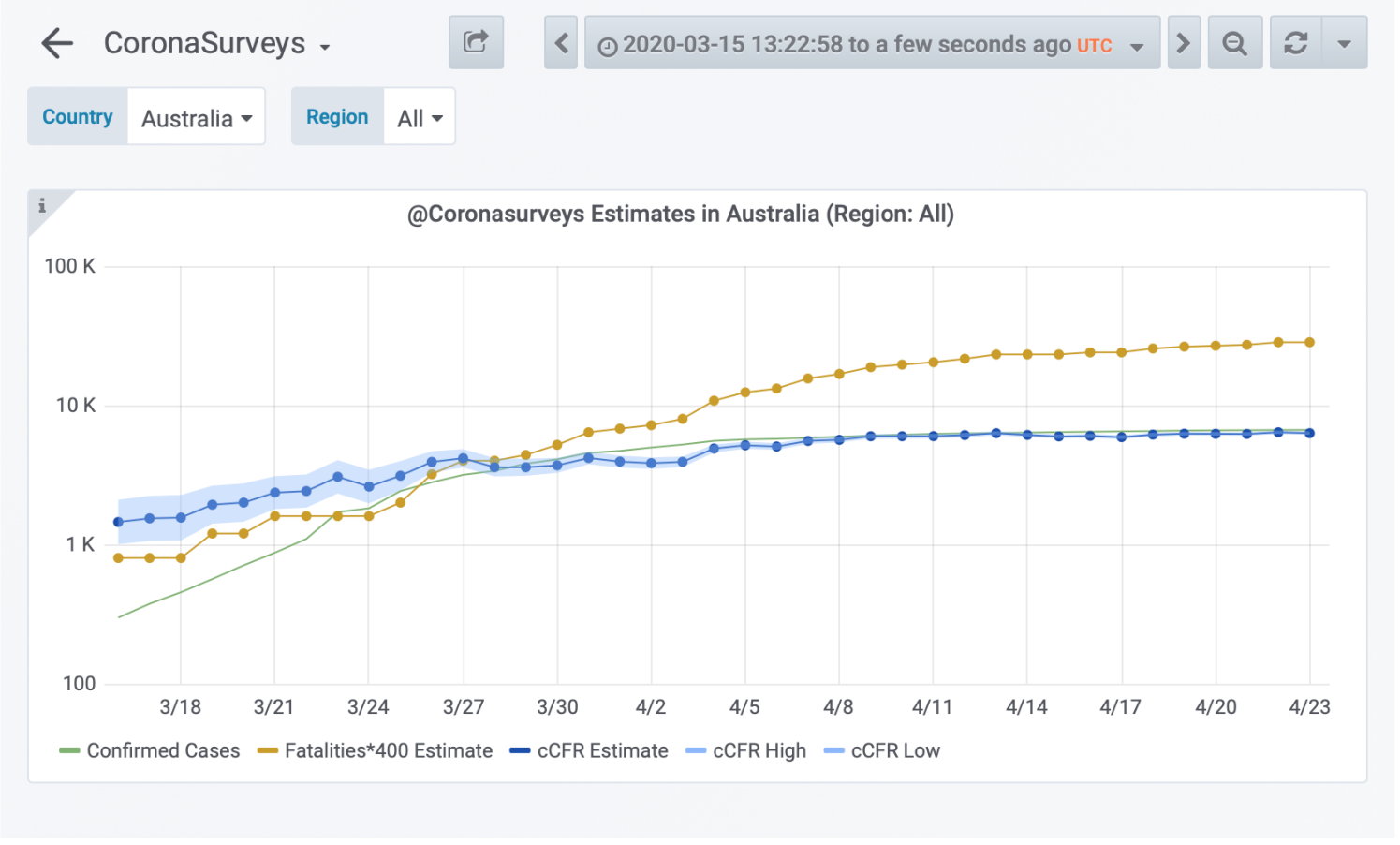


# Portfolio

## France



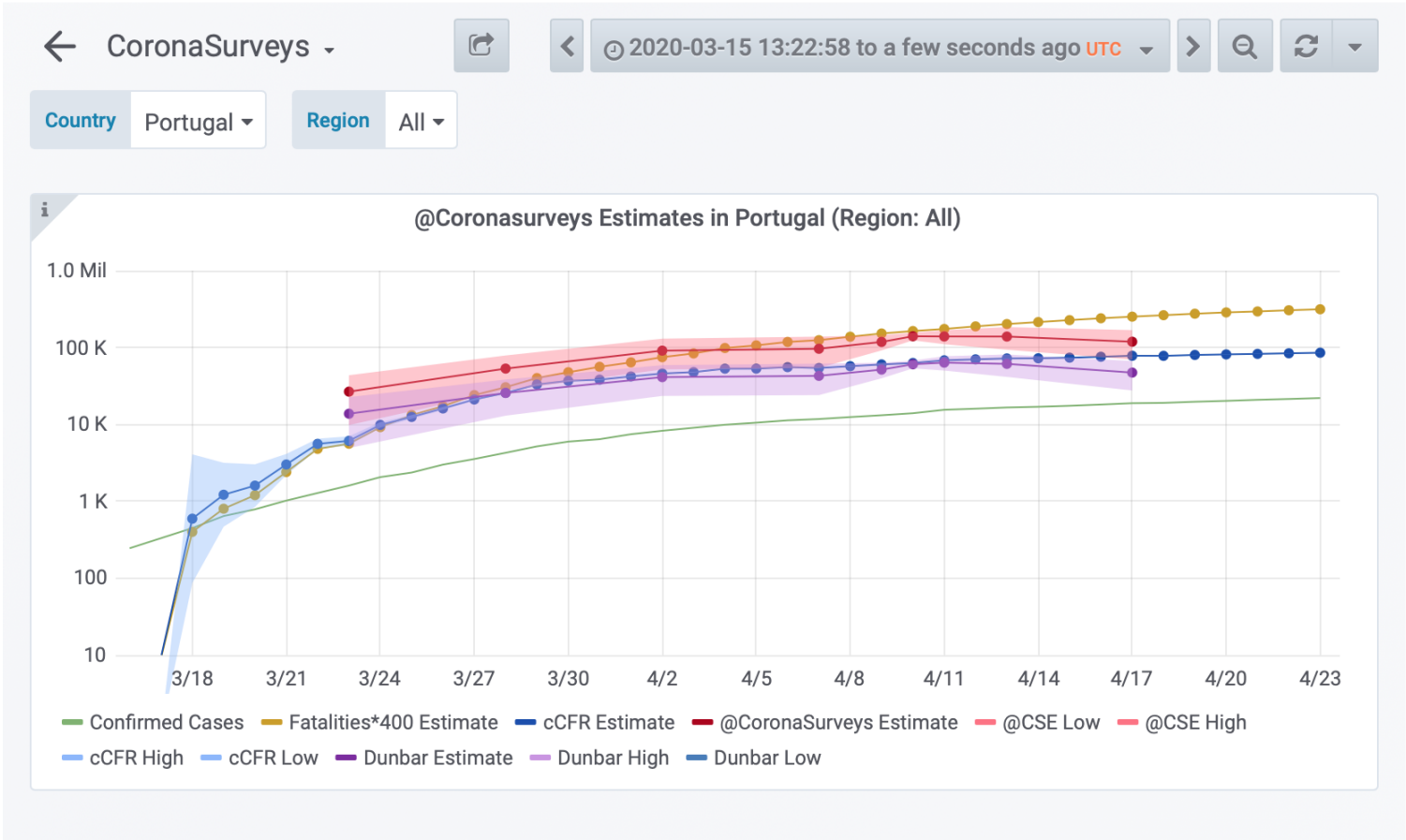
# Portfolio
## Australia




Closely matching Wuhan baseline

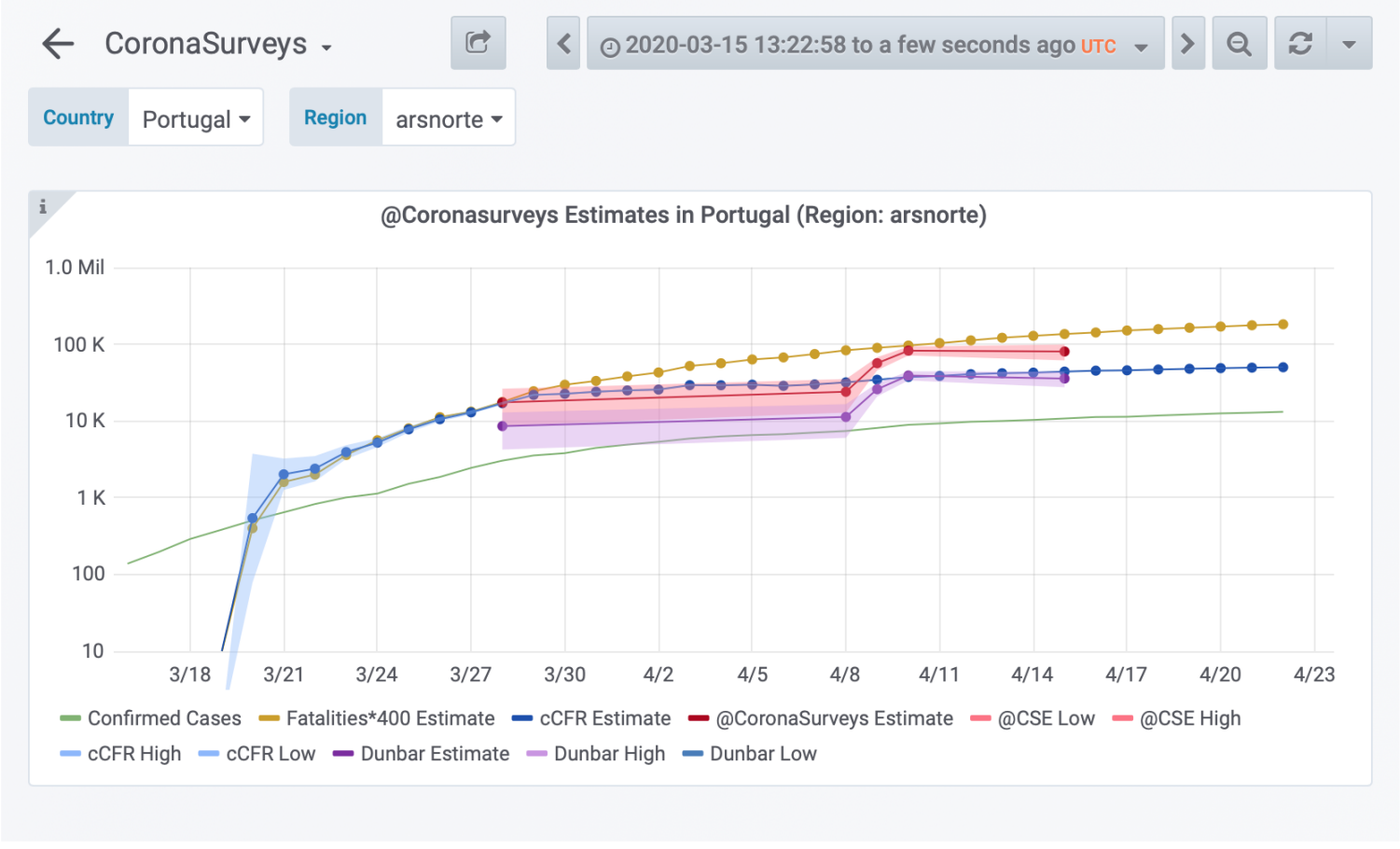
# Portfolio

## Portugal



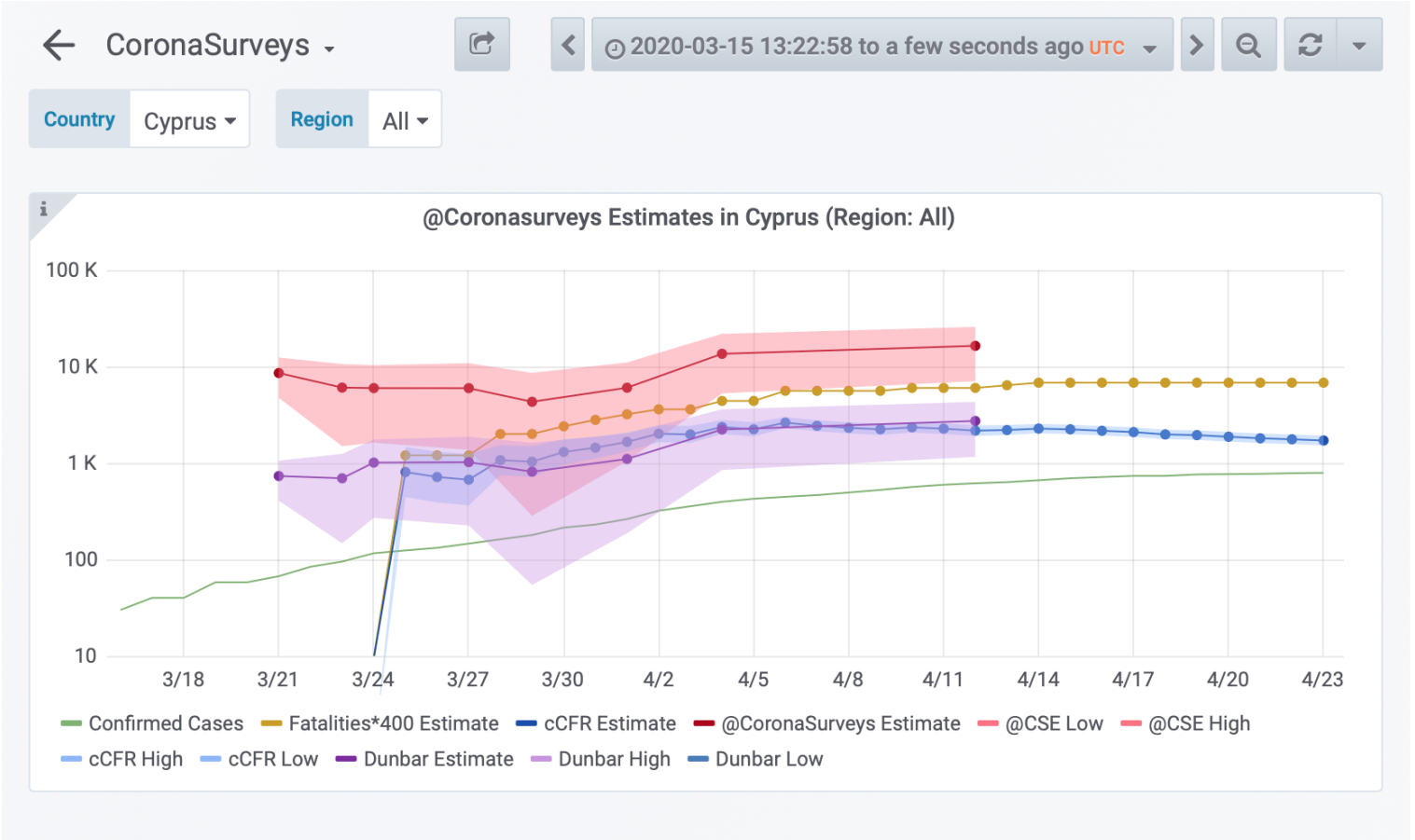
# Portfolio

## Portugal ARS Norte



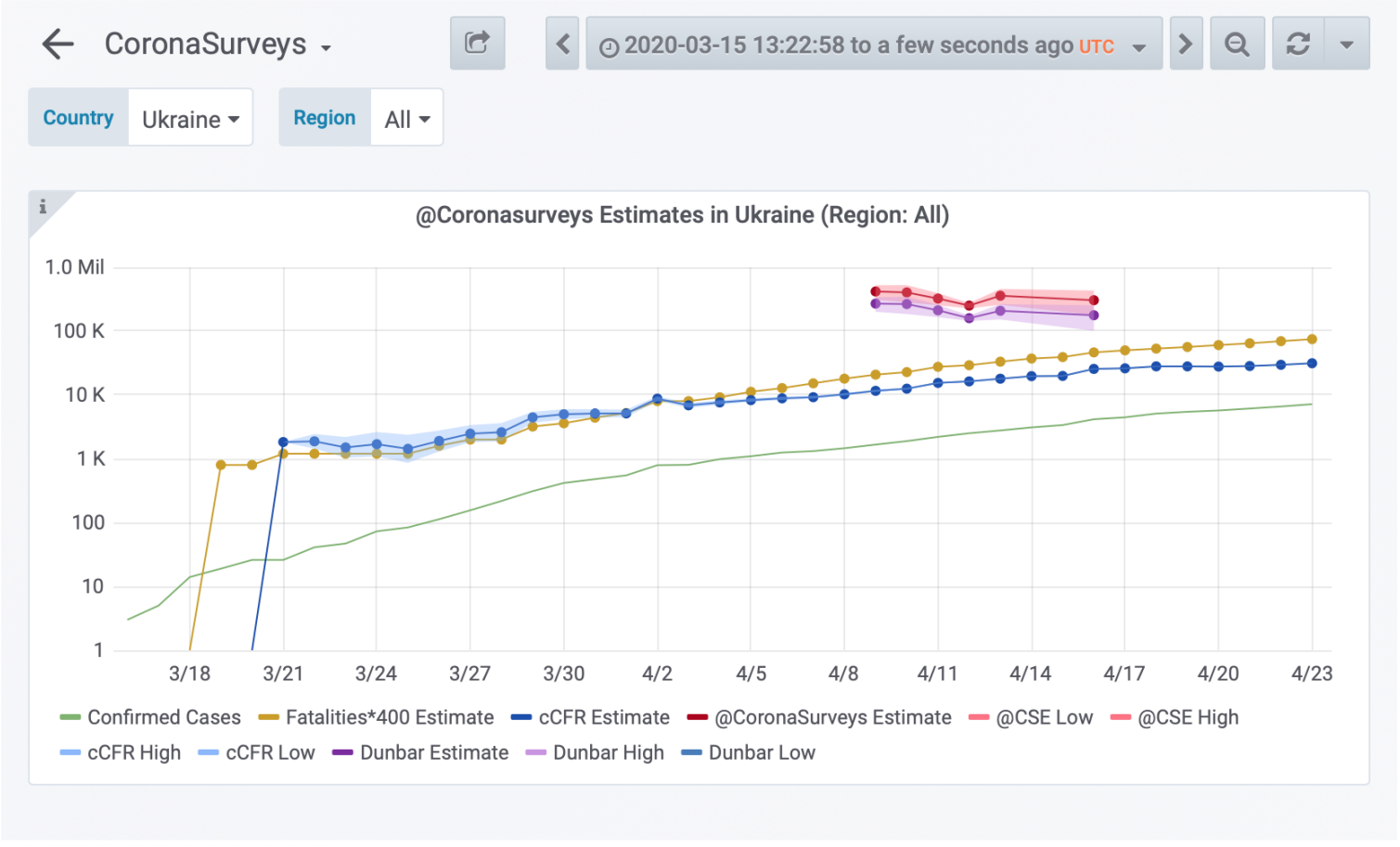

Data by regions in some countries

# Portfolio
## Cyprus




Survey estimates preceded fatalities derived estimates

# Portfolio
## Ukraine




Potential under-reporting of deaths

# Questions?

@CoronaSurveys at http://coronasurveys.org


Both our raw and processed data is openly available